MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2018/2019 SESSION

TDS 3301 – DATA MINING

(All Sections / Groups)

15 MAR 2019
3:00 p.m – 5:00 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 4 printed pages including cover page with 4 questions only.

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of marks for each question is given.

3. Please write all your answers in the Answer Booklet provided

Question 1

a) Compare deep learning with traditional machine learning algorithms in the aspect of (i) data dependencies (ii) feature engineering and (iii) interpretability. [3 marks]
b) State four reasons that drive the popularity of deep learning? [4 marks]
c) Why deep representations are used to solve complex problems i.e., image classification and speech recognition? [3 marks]

Question 2

	Male	Female	Total
Fiction	400	350	750
Non-fiction	200	50	250
Total	600	400	1000

A contingency matrix is as shown above. The data analyst wants to investigate whether or not gender is associated with reading fictions. He forms the two association rules as follows. The minimum support and confidence required are 40% and 60%, respectively
Male →Fiction
Female → Fiction

a) Calculate the support and confidence for the two association rules. Show your steps. [2 marks]
b) What can you conclude about the rules from your answers in (a)? [2 marks]
c) Calculate lift for the two rules. [2 marks]
d) What can you conclude about the rules from your answer in (c)? [2 marks]
e) Suggest two measures to use if the data contain a large number of null transactions. [2 marks]

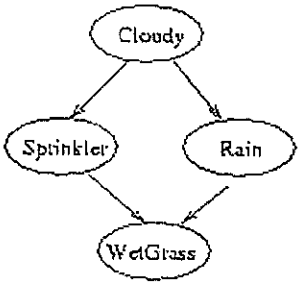
Continued...

Question 3

The following questions are related to a Bayesian Network.

a) What is a Bayesian Network? [2 marks]
b) Why a Bayesian Network is a directed acyclic graph? [1 mark]
c) Explain the qualitative and quantitative components of a Bayesian Network. [2 marks]
d)

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S R	P(W=F)	P(W=T)
F F	1.0	0.0
T F	0.1	0.9
F T	0.1	0.9
T T	0.01	0.99

For the above Bayesian Network,
(i) Write the chain rule to compute the probability for all the nodes, P(Cloudy,Sprinklet,Rain,WetGrass). [2 marks]
(ii) Simplifying (i) by considering conditional independence relationships. [2 marks]
(iii) What is the fundamental assumption made for writing (ii)? [1 mark]

Continued...

Question 4

Suggest a method to solve or mitigate each of the following problems. Explain why you think the methods you suggest may work.

a) There are missing values for data attributes, i.e., salary and education spending. [2 marks]

b) You plan to use a classification algorithm that is simple in calculation and supports only nominal class values. But your large data set contains numerical class attributes. [2 marks]

c) The data set you use contains two classes that are very unbalanced in distribution (let's say 95% vs 5%). The minority class gives a very low true positive rate during classification. [2 marks]

d) When you visualise your data set, you find out that the classes are overlapped together. This causes low classification accuracies for the classes. [2 marks]

e) You had tried many single classification algorithms on the small data set X, which is balanced in its class distribution, but you always get unsatisfactory accuracies. [2 marks]

Formulae:

Support, the probability that a transaction contains $X \cup Y$

Confidence = $\sup(X \cup Y) / \sup(X)$

Lift = $\sup(X \cup Y) / [\sup(X) * \sup(Y)]$

The chain rule for probability

$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\ P(b \mid c, .. z)\ P(c \mid .. z)..P(z)$

The chain rule of probability considering conditional independence relationships

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(Y_i))$$

End of Page